Generative AI in Medicine

Divya Shanmugam[1], Monica Agrawal[2,3], Rajiv Movva[1], Irene Y. Chen[4,5], Marzyeh Ghassemi[6], Maia Jacobs[7,8], and Emma Pierson[1,9]

[1]Department of Computer Science, Cornell Tech, New York, United States, 10044
[2]Department of Biostatistics and Bioinformatics, Duke University, Durham, United States, 27705
[3]Department of Computer Science, Duke University, Durham, United States, 27708
[4]Department of Computational Precision Health, UC Berkeley and UCSF, Berkeley, United States, 94709
[5]Department of Electrical Engineering and Computer Science / Berkeley AI Research, Berkeley, United States, 94709
[6]Department of Electrical Engineering and Computer Science/Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, United States, 02139
[7]Department of Computer Science, Northwestern University, Evanston, United States, 60208
[8]Department of Preventive Medicine, Northwestern University, Evanston, United States, 60208
[9]Department of Population Health Sciences, Weill Cornell Medical College, New York, United States, 10044. Email: emma.pierson@cornell.edu

Abstract

The increased capabilities of generative AI have dramatically expanded its possible use cases in medicine. We provide a comprehensive overview of generative AI use cases for clinicians, patients, clinical trial organizers, researchers, and trainees. We then discuss the many challenges – including maintaining privacy and security, improving transparency and interpretability, upholding equity, and rigorously evaluating models – which must be overcome to realize this potential, and the open research directions they give rise to.

1. Introduction

Excitement about the promise of generative AI in medicine has inspired an explosion of new applications. Generative models have the potential to change how care is delivered (1–5), the roles and responsibilities of care providers (6, 7), and the communication pathways between patients and providers (8, 9). Further upstream, generative models have shown promise in improving scientific discovery in medicine (through both clinical trials (10, 11) and observational research (12, 13)) and facilitating medical education (8, 14). These developments are a direct result of technical advances in generative AI, which have drastically increased the ability to generate realistic language and images, and raise important questions about how to integrate generative models into medicine.

Generative AI is the latest in a series of technical advances that have driven major shifts in medicine. Past significant advances include the adoption of electronic health records (EHRs); the integration of robotics into telesurgeries (15); and the incorporation of predictive models and continuous monitoring as foundational infrastructure for new diagnostic tools (16, 17). But the introduction of new technologies into health settings inevitably introduces new challenges to overcome. For instance, the introduction of EHRs led to increases in data privacy concerns and data security breaches (18, 19). And while the introduction of EHRs has led to significant reductions in medical errors and improvements in medical guideline adherence overall (20), they have also introduced other types of errors (21). Similarly, continuous monitoring devices in healthcare settings have resulted in pervasive alert fatigue (22). Overall, the integration of technologies into medicine requires an iterative design process which addresses pitfalls and amplifies benefits (23).

So, too, with generative AI. As generative models become a leading area of research and deployment in medicine, we provide a comprehensive review of the new applications they enable and the new challenges they create, with a particular focus on how users could interact with generative models. We first provide a brief overview of generative AI, detailing salient types of generative AI and how they fit into the broader landscape of machine learning in medicine. We next discuss the myriad use cases for generative AI in medicine, organized by potential users: clinicians (§2.1), patients (§2.2), trial organizers (§2.3), researchers (§2.4), and trainees (§2.5). We then highlight the challenges (§3) that must be addressed to realize this potential and safely deploy generative models (including ensuring informed consent, protecting privacy, and improving transparency, among many other considerations) and discuss future directions for research throughout.

1.1. Background on Generative AI

Generative modeling is a fundamental AI paradigm which stands in contrast to *predictive modeling* (also called discriminative modeling): predictive models are given an input and seek to predict its label but do not attempt to model the input, whereas generative models do seek to model the input. For example, while a predictive model might be given a clinical note (the input) and try to predict whether the note indicates the presence of cancer (the label), a generative model would aim to model the distribution of clinical note text itself. The fact that generative models are trained to model the entire data distribution affords them the powerful ability to *generate new data*: for example, to write new clinical notes.

The basic generative modeling paradigm far predates the current surge of interest in generative AI. For example, classical generative modeling methods such as Markov chains have been used to model sequences of words for decades (24), and in theory could be used

to write clinical notes. In practice, however, classical generative modeling methods do not generate sufficiently realistic content to be useful, especially on complex medical data. The current surge of interest is fueled by a drastic increase in generative modeling capabilities, which has been driven by scaling models to larger deep learning architectures and larger datasets (25). These improvements have, as we describe, expanded the number of useful applications of generative AI models, and have spurred interest in applying these models to domains outside of core machine learning (26).

We summarize three categories of generative models organized by the *type of data* on which the model operates: (1) text, (2) images, or (3) text and images. For each category, we focus on the state-of-the-art models that are currently in use. While we limit our discussion to text and image data since these are most relevant to the use cases we subsequently discuss, generative models for other types of data (for example, physiological signals and molecular graphs) are an emerging area of clinical AI research (27–29). For more comprehensive overviews, we refer the reader to (30).

For modeling text, **large language models (LLMs)** are the dominant approach, with substantial performance gains in recent years. LLMs commonly use the transformer neural network architecture (31) to perform next word prediction: given a sequence of words, what is the most likely word that will come next? That is, for a context sequence $x_1, \cdots, x_n$, an LLM is trained to predict $p(x_{n+1} \mid x_1, \cdots, x_n)$ (32). What makes an LLM "large" is the size of its deep learning architecture, and the amount of data and compute used to train it; most language models in use today are considered LLMs. Training an LLM usually consists of three phases: first, the LLM is *pre-trained* on a large corpus of text scraped from the Internet; second, it is fine-tuned on *instruction-following* examples, wherein user questions or instructions like "Convert this discharge report into layperson's terms" are followed by reasonable responses; third, the LLM is tailored with *human feedback*, where humans choose which of two possible responses they prefer to capture fine-grained preferences (33). Each of these phases can be tailored more specifically to medicine: some models are pre-trained on medical corpora, usually PubMed, in addition to or instead of the entire Internet (34, 35); some models are trained specifically to respond to medical questions using datasets like MedQA (36, 37); and there are emerging datasets of physician-written responses to medical queries to help align LLMs to medical best practice (38). While the use cases we will discuss primarily involve models trained on human language, models with similar transformer architectures can also be trained on other types of sequential biomedical data. For example, electronic health record models have been trained on sequences of ICD codes (39, 40); protein models have been trained on sequences of amino acids (41); and DNA models have been trained on nucleotide sequences (42, 43).

For modeling images, **diffusion models** (44) have recently become the method of choice, largely surpassing the previous generation's generative adversarial networks (45). Given an unlabeled training distribution of images, a diffusion model learns to generate new, synthetic images that look like images from the training distribution. To train a diffusion model, a real image x_0 is progressively corrupted to produce x_t, an image after t corruption steps that looks like random noise. The model is trained to reconstruct the original clean x_0 from noise x_t by learning the distribution $p(x_{i-1} \mid x_i)$. A well-trained diffusion model can then start from randomly-sampled noise, and produce a new image that is not in the training set but appears to have been drawn from the same distribution. Medical diffusion models have been trained on several different image types, such as chest X-rays, dermatoscopic images, and pathology slides (46). To improve the biological validity

of generated images, datasets and methods are being developed to fine-tune models using physician feedback (47). Synthetic image generation can be a useful augmentation technique in data-constrained settings, where supervised ML models may benefit from synthetic datapoints; recent evidence suggests this technique can help improve model robustness for pathology and radiology tasks (13).

For tasks involving text *and* images, there are two key types of generative models: **text-to-image (T2I)** and **vision-language models** (VLMs)[1]. T2I models accept a piece of *text* as input, which is used by a text-conditioned diffusion model to generate a corresponding image as output. These models consist of two components: a text encoder model (i.e., a transformer (48)), and a diffusion model which generates the image using the text encoding. T2I models are pre-trained using general image caption datasets; they can then be fine-tuned for medicine, e.g. using chest X-rays and corresponding radiology reports (49). T2I models further expand the possibilities of synthetic images, for example by allowing researchers to generate training data for a specific patient pathology. Relatedly, VLMs take in an *image* as input and generate text involving the image as output (50). VLMs consist of an image encoder model (e.g., a convolutional neural network or a vision transformer (48, 50)) and a large language model which can generate text based on the image encoding. VLMs also require large image-text datasets, which can include image captions or reports, but also answers to visual questions, like "Does this chest X-ray exhibit pleural effusion?" (51). They can be applied to tasks like question answering and report generation for pathology or radiology (52, 53).

Each of these model classes have natural applications in medicine. Many clinical processes and decisions involve unstructured text (in the form of clinical notes, online health information, and treatment plans) and medical images. Moreover, images and text often appear in combination, as is the case with radiology reports. The next section elaborates on the potential to combine these generative modeling paradigms with existing data and processes in medicine.

2. Use cases for generative AI in medicine

The use cases for generative AI in medicine are numerous (Figure 1). We organize our discussion around key constituents in patient care: clinicians, patients, clinical trial organizers, researchers, and trainees. Within each role, we highlight core responsibilities that can be transformed by the introduction of generative AI.

2.1. Clinicians

Generative models have the potential to improve clinicians' ability to provide care in multiple ways: by assisting with writing and documentation (reducing administrative burden and physician burnout, which are major concerns (54)); assisting with diagnosis; retrieving patient data; and supporting evidence-based medicine. Each use case suggests that generative models can act as a useful tool for delivering care personalized to patient needs.

Assistance in writing The amount of documentation in the electronic health record is a leading cause of physician burnout (55). Outside of their shifts, clinicians frequently spend

---

[1]The latter are sometimes also referred to as image-to-text models.

significant time finishing clinical notes and communicating with patients; generative AI provides an opportunity to speed up both of these documentation processes.

The earliest deployments of generative AI-driven writing assistance have been in ambient documentation, in which recordings of patient-provider communications are used to generate initial drafts of notes (6). In initial pilots, providers and patients alike have indicated that the technology facilitates improved interactions, and providers have noted a decrease in time spent completing notes after their shifts.

In another example, the past few years have seen an uptick in provider-patient messaging through the electronic health record. While this has provided a straightforward path for patients to surface concerns and questions to their providers, providers have had to similarly spend increasing amounts of time responding (55). There have already been several pilots using large language models to respond to patient queries (1, 3). Additional work has explored having large language models prompt *patients* with follow-up questions automatically so that clinicians have full patient context with fewer back-and-forths (9). While clinicians report high usability of response drafting tools and decreased burnout, early results show that significant editing is still required, and effects on documentation time have yet to be seen.

Across both of these use cases, there are significant nuances which must be addressed for successful deployments. First, the clinical note writing process is not one-size-fits-all, as workflows differ significantly between specialties and clinicians. Second, significant effort is necessary to ensure that automation in the writing process does not lead to automation bias or decreased agency from clinicians (56). Clinical notes often serve not only as documentation for future reference, but also as an active space for iteration and re-analysis (57) as, in many specialties, notes are not written at once, but in several sessions stretched over a patient's stay (58).

Assistance in diagnosis Several works have assessed the efficacy of using generative models to provide diagnoses from patient information (e.g., medical imaging and lab test results). While current generative models do not identify the correct diagnosis with sufficient reliability to be used without clinician supervision (59–61), their generated differentials (i.e. sets of diagnoses) often contain the correct diagnosis and could be useful as a tool to *augment* clinicians by expanding the set of diagnoses a clinician considers (62). Such a tool could be particularly useful in contexts where identifying candidate diagnoses is difficult (for example, rare diseases (63) or challenging clinical cases (64)). (61) find that a large language model includes the correct diagnosis in its differential in 88% of cases, nearing the diagnostic performance of clinicians, who include the correct diagnosis in 96% of cases. Other works have studied the accuracy of generative model diagnosis given both images and text (65–68); (65), for example, demonstrate how a vision-language model can identify the correct skin condition, given an image and a text prompt, in 80% of cases. Overall, diagnostic performance depends on both the task and generative model (69, 65, 70), and while significant progress must be made for generative models to provide diagnostic assistance to clinicians, the initial results are promising. Ultimately, such a tool could be used in service of precision medicine by providing diagnostic assistance personalized to a patient's medical history.

Information retrieval of patient data Clinical practice requires a significant amount of *synthesis* of a patient's past health narrative (e.g., understanding past medications, lab

trends, and diagnoses) with their current clinical state (57). In current EHRs, clinicians often have trouble finding the relevant information needed for contextualized clinical care, both due to the fragmentation of EHRs and the amount of information trapped in free-text notes (71). Finding relevant labs or medical history can often involve time-consuming and disjointed navigation through different sections of the electronic health record. Retrieval-augmented generation (RAG), in which generative models retrieve information from an external database and use it to inform the language they generate, provides a possible avenue for clinicians to conveniently surface data in a single unified process, powered by natural language (2). RAG combines the flexibility of generative AI with more classical information retrieval systems; for example, a clinician could query a generative model with "relevant family history for chest pain" or the "result of the patient's last colonoscopy" to retrieve information across a patient's EHR. Going even further, the model could proactively surface information likely to be relevant, based on the current stage of a visit. Here, open questions remain on what information to automatically surface, when to do so, and how to best display it (72, 58). Previously, researchers have found it useful to draw on past interaction patterns in the electronic health record in order to inform interface design (73).

Evidence-based medicine Evidence-based medicine requires bringing new findings from clinical research to the bedside, but it can be nearly impossible for clinicians to keep up with the pace of clinical research (74, 75). Large language models can make it easier to organize and query clinical trial information at the point-of-care. Existing explorations have included NLP-generated databases of cleaner, searchable clinical trial information and natural language interfaces for interacting with clinical guidelines more directly (7, 76, 77).

2.2. Patients

Patients often express a desire for increased involvement in their care, e.g. in the form of shared decision making or increased access to information (78). Below, we describe a few ways generative AI could impact the patient experience. We emphasize to the machine learning community the necessity to embrace *participatory design*, where proposed tools or interfaces (for example, interfaces to surface health information) are built in collaboration with patients and center stakeholder perspectives, as has been done in informatics and human-computer interaction (79–82).

Searching for Online Health Information Patients often leverage the Internet to help answer their health questions, particularly when they have limited access to care (83). For those with access, benefits have included increased ability to participate in decision making, more informative questions during episodes of care, and improved adherence to instructions (84). Recent surveys show that generative AI has already seen significant adoption for information seeking between appointments (85, 86). Unlike existing search engines, generative AI enables patients to pose more specific queries and ask follow up questions, allowing for more tailored responses and conversational searching where questions and responses can build upon one another. However, the current generation of models can provide plausible but inaccurate information, making it challenging for patients to discern when to trust model outputs (87). Patients are still faced with the challenge of discerning if the information is accurate, and current generative AI chatbots provide few references. We expand on these issues in later sections.

Increasing Patient Engagement Increased patient engagement can enable patients to better understand their own health conditions and care plan. This engagement can also lead to an increase in patient-reported outcomes (e.g., self-tracking and sharing of symptom burden); this in turn can enable clinicians to better understand their patients' conditions in between visits (88). One approach to increased patient engagement is the use of patient portals; however, utilization of patient portals remains low (89). Given their potential, there have been several long-running suggestions for how to increase utility of patient portals that are now more feasible with generative AI (90–92). As an example, clinical notes contain valuable information for patients but were not written with patients as the intended audience: they are filled with jargon, and patients cite the difficulty of medical jargon as a major barrier to comprehension (93). Recent work has explored using generative models to *translate* clinical notes into patient-friendly language and visualizations, with the opportunity to personalize to information needs of patients and how they want that information presented (4, 5, 94). The opportunity of patient-friendly text simplification extends past clinical notes alone to other facets of health literacy, e.g., medical literature and patient consent forms (95, 96).

2.3. Trial Organizers

Clinical trials provide critical evidence to update and improve clinical practice. Conducting these trials, however, remains challenging: only 20% of clinical trials in the United States complete within the planned timeframe (97), and of those that do, only a small fraction are published (98, 99). The difficulty of clinical trial design is in part due to the complexity of interaction and documentation involved; trials can fail due to incorrectly designed protocols, insufficient participant registration, or high patient dropout (100, 101). Generative interfaces present the potential to rework key components of this pipeline.

Protocol Design Clinical trial design begins with the creation of a protocol, which collates existing research, study aims, and regulatory requirements into concrete steps detailing how the trial will proceed. Protocol creation requires significant manual effort (102), and existing work has illustrated the value in using generative models (specifically, large language models) to expedite the process (10, 11, 102–105). Several works center on the generation and evaluation of exclusion and inclusion criteria (106, 103, 10, 105), while others propose the use of large language models to retrieve relevant past clinical trials to inform the construction of a new protocol (11, 102). (104) employs large language models to evaluate protocols for bias automatically, while (10) uses large language models to generate trial inclusion and exclusion criteria based on details of the setup expressed in natural language. Together, the literature thus far highlights the potential for generative interfaces to reduce the time required to construct a successful protocol.

Participant Recruitment & Retention Equipped with a protocol, clinical trial designers must recruit a sufficient number of participants to conduct the trial. Generative interfaces can be used to identify suitable trial participants by parsing both the criteria and patient history (107–110). Early results suggest that large language models can reduce the number of eligibility criteria a clinician must manually check to assess eligibility by 90% (107) and reduce the time it takes to assess eligibility by 42% (109).

As the trial progresses, patient dropout can threaten its validity and success. Patients drop out of clinical trials for a multitude of reasons, one of which is poor communication

with trial recruiters and clinicians (111, 112). Patient dropout is particularly salient to decentralized clinical trials, which are conducted at non-traditional sites (e.g. a patient's home, or a local clinic) and are known to recruit more diverse patient populations compared to their standard counterparts (113). (114) propose the use of AI to improve patient engagement in these settings, which could include chat interfaces to answer trial-related patient questions as they arise. Such an interface could effectively and efficiently address patient concerns and misconceptions about a trial, including the potential for adverse events (115) or the importance of trial participation regardless of treatment outcome (112).

2.4. Researchers

Designing a useful medical study is a time-consuming process. Researchers comb through large bodies of literature, across multiple disciplines, to identify open questions and understand the status quo in different application areas. Later stages of research, including dataset construction, hypothesis generation, and code generation are no easier. Here, we highlight how generative models can help alleviate significant manual effort at each step.

Literature review Efforts to collate existing literature into a coherent research question precede any effort to execute the study. Large language models have been shown to be a promising tool for problem generation through automated systematic literature reviews (12, 116). Here, a generative interface can allow a researcher to automatically assemble a clinical review by querying thousands of clinical abstracts, substantially reducing the effort required to perform a systematic review. (12) demonstrate how a popular large language model (GPT-4) can identify relevant papers at 91% accuracy compared to human evaluation, with the ability to justify the inclusion and exclusion of particular papers.

Dataset construction Generative AI can improve the quality, quantity, and diversity of datasets in medicine (117, 13, 118). To produce such datasets, generative models can be used in two ways: to *generate completely synthetic data*, or to *extract structured information* from existing unstructured data.

Using generative models to create synthetic data has shown promise in compensating for gaps in existing datasets (13, 119, 120). (119) show that synthetically generated images can be used to improve a machine learning model's ability to detect COVID-19. Beyond improvements to accuracy, (13) illustrated how augmenting training data using generative models can improve the fairness of the resulting diagnostic classifier. These findings hold in the context of natural language; (120) demonstrate how diagnostic classifiers trained on generated text perform comparably to those trained on real datasets of the same size, suggesting that synthetic data is a promising approach to addressing data limitations.

Generative AI can also be used to extract structured information from semi-structured data or to label existing data, across both natural language and imaging. Research on health equity, for example, relies on structured fields describing patient demographics to assess the severity of health disparities. Large language models can be used to extract demographic data from unstructured text (e.g. clinical notes) (121), thus enabling comparisons of health outcomes between demographic groups. Similarly, several generative AI tools have also been developed to measure morphological features from large cohorts of histopathology images using natural language prompts (122–124). Generative models can also be used to alleviate the burden of data annotations by providing synthetic labels (i.e. predictions of

ground truth) for unlabeled examples (125); for example, one could use a generative model to suggest candidate segmentations of medical images (126).

Hypothesis generation Given a dataset, generative models can be used to surface hypotheses (127–130). (127) examine the use of generative models to produce natural language hypotheses (i.e. "customers tend to buy shoes that match the color of their shirt"). They find that the resulting hypotheses confirm expected relationships, provide new insights, and outperforms supervised baselines. (128) frame the discovery of drug-specific side effects as a task for a generative model; given patient feedback for different drugs, the generative model is tasked with describing differences between drug-specific patient feedback in natural language.

Code generation Generative interfaces can also be used to write code based on natural language prompts, which could lower the barrier for researchers to perform quantitative analysis of large-scale datasets (131). (131) demonstrate that GPT-4 is capable of autonomously producing code to train models for disease screening and diagnosis. Collaboration with a code assistant has been shown to improve programming productivity (132), and could help facilitate quantitative analysis of increasingly large observational health datasets.

2.5. Trainees

Generative interfaces are already in widespread use by medical trainees (133). The rapid uptake of these tools among trainees suggests the potential for generative interfaces to significantly transform medical education. Two promising applications are the use of generative interfaces to create practice clinical scenarios and provide feedback that targets student-specific areas for improvement.

Case creation Designing realistic clinical scenarios to test understanding is a critical yet difficult task. Those in charge of clinical curriculum design could use large language models to generate compelling multiple choice questions, as (134) have demonstrated in the context of surgical education. Doing so could also address the lack of diversity in clinical vignettes (135), and allow the generation of problems that better reflect patient populations trainees are likely to interact with (14, 136, 137). (14), for example, develop a tool to use large language models to produce 30 distinct cases in under an hour (including manual human review). Simulated patients could also be used to simulate real-world interaction. For example, the process of collecting patient history could be taught through a generative interface, in which a large language model responds to a medical student's requests for information based on a synthetic patient profile. (138) have shown that such a set-up is well-received by medical students, and that more than 97% of the generated answers were deemed clinically plausible. Each of these uses cases presents an opportunity to reduce the resources required to train medical students to deliver care.

Providing personalized feedback Generative interfaces can also provide tailored feedback to students (139). (139) show that large language models can provide students with more coherent, process-oriented feedback compared to human instructors, while retaining high agreement with human-generated feedback. There is significant opportunity to apply these findings in medicine; for example, one could use large language models to provide feedback

on a medical student's efforts to communicate health information to patients. Indeed, (8) showed that when generative models are used to provide feedback on clinical notes, the resulting notes are more complete, concise, and correct across four distinct specialties.

Feedback for trainees can take many forms; in surgery, for example, non-generative AI is already being used to provide automated assessments of surgical procedures in simulated environments (140), which helps trainees safely gain familiarity with procedures. (141) develop a generative model to output the optimal surgical path and show that real-time guidance leads to significant improvements along multiple surgery-specific performance metrics, including the number of attempts required to complete the surgery and risk of tissue damage.

3. Challenges and directions for future work

Fully realizing the opportunities to apply generative AI in healthcare requires significant progress on a number of challenges we describe below, and illustrate in Figure 2. Generative AI interfaces are far from perfect, and we are only beginning to understand the impacts of such interfaces on clinical decision-making. We have already seen examples of the potential harms such interfaces have caused that warrant attention. For example, (70) demonstrate how large language models fail to follow diagnostic guidelines up to 36% of the time in an evaluation across realistic patient cases. Below, we enumerate challenges we foresee as critical to address as the intersection between generative interfaces and healthcare evolves, and discuss future directions for research throughout.

3.1. Ensuring informed consent

Informed consent is a foundational principle of medical ethics which states that a patient must have access to sufficient information about a medical procedure (including risks, benefits, and alternatives) (142–145). Achieving informed consent when using AI models raises new challenges which are a topic of active discussion (146): for example, how do we provide patients with comprehensive, accurate, and understandable information about complex models whose behavior is not fully understood even by their own creators (let alone the clinicians using them)? How do we ensure patients are consenting to the use of their data if it is used in model training? These issues similarly apply to generative models, which also raise new challenges (147, 148). For example, when asked for their concerns about the use of generative AI models to transcribe and summarize patient-clinician conversations, providers expressed concerns about whether patients could meaningfully consent to the collection of this data (147). Similarly, the use of generative models in chatbots that interact with patients raises new concerns about informed consent (148). When chatbots and other generative AI tools are implemented in care practices, patients need to be given the option to decline the use of the models in their care and the use of their data. Patients must be provided with clear information that they are interacting with a chatbot, who the chatbot's creators are, and the uses and limitations of the technology.

Contrasting with these concerns, generative models also show promise for *improving* the informed consent process (149) by making consent forms easier to understand. A study comparing LLM-generated consent forms to those created by five surgeons for six common medical procedures found that the LLM-created forms were more readable and accurate than those created by surgeons. In this way, LLMs can help advance equity in

medicine, through the creation of consent forms that are more accessible to a broader audience. Current medical consent forms are often written at a high reading level and describe complex procedures (150). As a result, many people, especially non-native English speakers and individuals with low literacy, are at increased risk of consenting to research and medical procedures without being fully informed (151). While LLMs offer promise in improving consent documentation readability, more work is needed to ensure consent documents developed by LLMs contain comprehensive information (152).

3.2. Maintaining privacy and security

The use of generative models in medicine raises substantial privacy and security challenges (153), given the sensitivity and legal protections of medical data. One challenge is that generative models perform best (and are more likely to generalize) when trained on large, multi-institutional datasets, raising the question of how to share data across multiple institutions in a privacy-preserving way. Technical approaches like federated learning (154–156) offer one approach to this, although recent work has indicated that privacy violations are still possible in this setting (157–159). The creation of large de-identified datasets which can be securely shared with researchers (160, 161) is also an important catalyst for generative model research which institutions and policymakers should facilitate.

After training generative models, a second challenge is *deploying* models trained on sensitive data in a secure and private way. Past work has demonstrated that generative models can leak sensitive data by memorizing their training datasets and revealing private data in response to adversarially crafted prompts (162–166). Past work also suggests that these problems may grow worse as models continue to scale because larger models possess greater capacity to memorize the training data (164). Mitigating these challenges remains an active area of research which is essential for safely deploying models trained on sensitive health data.

Finally, and more optimistically, generative AI also holds potential for sharing data in a more privacy-preserving way, via generation of synthetic datasets which mimic properties of a real dataset while preserving patient privacy (167–169) (see §2.4 for further discussion of synthetic datasets). For example, (167) demonstrate how synthetic patient records produced by a generative model can be substituted for real data at no loss of performance, with significant improvements to patient privacy.

3.3. Improving transparency and interpretability

Modern AI models are opaque for a number of reasons, and generative models are no exception. A first major challenge is a lack of *transparency*: basic details of generative models are often not disclosed, including training data, training methods, model architecture, capabilities, limitations, and risks (170). Lack of transparency causes several harms (170): it makes it more challenging for policymakers to regulate models; for users to assess when they will perform reliably; and for researchers to innovate on them. Consolidation around a small number of closed models risks heightening the lack of transparency (26). The sensitivity of health datasets, which often cannot be publicly released, also makes achieving transparency more challenging. A 2023 review of widely used generative models scored them on 100 granular transparency indicators and found they averaged only 37 out of 100 (170), though the average score had improved to 58 out of 100 when the review was conducted in May 2024 (171), suggesting that transparency can be improved and that systematic reviews

of the ecosystem are helpful.

A related, but distinct, challenge is *interpretability*: even if all details of a model are fully disclosed, understanding *why* the model gives the output it does can be extremely difficult. Without understanding why a model produces a particular output, it is difficult to know whether to trust the model, and when it will fail. For example, healthcare models have been known to rely on spurious features to make predictions, and without knowing what features a model is using, these failure modes are difficult to identify (172–174). Interpretability challenges are not unique to generative models, but occur with many modern AI models, including other deep learning architectures (175, 176). In general, interpretability methods (also known as explainable AI methods) have seen mixed success (177, 178); different methods can yield very different answers, and those answers may be misleading.

Similar interpretability challenges occur in the context of modern generative models, which can have billions or trillions of parameters, encoding non-linear, highly complex functions of the input data which are extremely challenging to understand or describe in a human-interpretable way (179). While language-based generative models can provide plausible-sounding explanations for their reasoning (180), seemingly improving interpretability, those explanations are not necessarily accurate (181). In general, the capabilities of generative models are currently advancing considerably more quickly than our ability to explain how they achieve those capabilities, which is concerning especially in high-stakes domains like healthcare.

Improving interpretability of generative models remains an active research area; proposed approaches include *local explanations*, which explain a single output from an generative model and *global explanations* which explain a model's behavior as a whole (180). In a qualitative analysis of local explanations for a vision-language model applied to pathology images, (182) find that the interpretations align with clinically known disease characteristics. Although the fidelity of such explanations is context-dependent, they remain a key ingredient in improving the transparency of generative models. Another recent line of work seeks to train generative models that are interpretable by design, by training models on paired images and text so the model can provide natural language annotations of generated images; (183) apply this approach to dermatology data, and find that the models can accurately annotate images, as verified by dermatologists. A final way to address interpretability challenges is simply *rigorous evaluation* of a model across a range of settings: even if it is not possible to understand exactly how a model produces its outputs, one can verify that they are reliably accurate.

3.4. Mitigating hallucinations

Recent work has shown that generative models sometimes output medical information that is incorrect (184) or hallucinated (185, 186). Hallucinated outputs in high-stakes medical settings can be dangerous: for example, they can harm patients without clinician access who rely on LLM-generated outputs for medical advice. Other work shows that LLM outputs can be hard to understand or non-actionable (187), which, while not directly harmful, may undermine widespread usability, especially for underserved patients.

One promising approach to reducing generative models' propensity to hallucinate is *retrieval-augmented generation* (RAG). As discussed above (§2.1), RAG integrates traditional approaches towards search and information retrieval into generative models: specifically, by retrieving information from a verified knowledge base to guide the text a language

model generates. RAG has been shown to reduce the extent to which large language models hallucinate (188) and can also make a model's information sources more transparent, increasing users' ability to assess their reliability. While RAG and other methodological developments continue to mitigate hallucinations (76, 189, 190), it's essential to continue validating accuracy and readability in each use case, especially more difficult or error-prone ones (e.g., as done in (191)). We further discuss challenges in evaluating LLMs below (§3.7). In addition to technical methods and audits, regulatory oversight of generative models will also help mitigate the harms of inaccuracies (192). Policymakers can, for example, encourage greater transparency in model development by mandating disclosure of adverse events and of important details which are currently often not disclosed, including training datasets, architectures, limitations, and biases.

3.5. Designing usable interfaces

Pioneers in human-computer interaction have called generative AI "the first new interaction model in more than 60 years" (193). When interacting with generative AI, users are now able to tell the computer their desired intent (e.g. create a summary of melanoma for patients that includes symptoms, treatment options, and management strategies), rather than the exact actions they want the computer to take (194). Thus, generative AI interactions can be efficient and low-burden (195), and offer a new way to design and develop novel health interventions (196). However, this new interaction paradigm brings usability challenges that have yet to be addressed. Numerous studies have shown that creating accurate and useful prompts for a generative AI platform is difficult for end users (197–201). Once a prompt is provided, end users then face the additional challenge of interpreting and evaluating the output (200, 202). As we have seen with previous AI technologies, when end-users are unable to accurately evaluate the output of a model, they can become overreliant or make erroneous decisions (203–206). To address these usability challenges, (207) recommend adopting user experience principles to guide the design of these systems. More work is needed to establish best practices for both the design and end-user evaluations of generative AI systems.

AI tools often face many obstacles to widespread adoption (208–211) leading to limited health impact (212, 213). Given similarities of generative interfaces with prior AI tools, we expect similar challenges to arise. First, healthcare professionals may be hesitant or skeptical about new generative AI interfaces, making them resistant to change. In studies which simulated clinical settings with patients, research has found that provider experience levels (214, 215), interface style (216), and time pressure (216) may all affect adoption likelihood. Second, training and education are crucial to ensuring that healthcare professionals can best leverage these new technologies, which can be costly and time-consuming. Exposure through formal education or prior experience with similar AI interfaces can make healthcare professionals more comfortable with AI tools, leading to a higher rate in adoption (217). Educating users on the strengths and weaknesses of commonly used generative interfaces has been shown to improve qualities of human-AI collaboration including accuracy (218) and reliability (219), and we expect these findings to hold true in medicine. Lastly, deployment of generative interfaces may require an initial investment in cost and resource allocation. While preliminary studies have shown that generative interfaces can save time — e.g., initial estimates show that it may save nurses around 30 seconds per generated message (220)— the potentially large upfront cost remains a key concern to active adoption.

3.6. Centering equity

Abundant previous work has demonstrated how biases in medical datasets can propagate into AI models (221–242). It is thus unsurprising that the use of generative models in medicine creates several equity-related challenges (243). Research indicates that larger models do not, on their own, necessarily resolve equity concerns; indeed, larger models have been shown to exhibit more covert forms of bias (i.e. prejudice against certain dialects) compared to smaller counterparts (244). (243) perform the largest-scale health equity evaluation of large language models to date, highlighting the complexity of equity challenges and the necessity of careful, multi-dimensional evaluations to identify and mitigate them.

A first challenge is mitigating stereotypes and bias in generated text. Like human clinicians (245), LLM-generated text has been shown to display medical stereotypes (246, 247). For example, (246) finds that when GPT-4 is asked to provide clinical vignettes about sarcoidosis, it generates vignettes about Black patients 97% of the time, exaggerating the true population skew. Due to these embedded biases, if patients specify their demographics when asking questions to LLM chatbots, there is a risk that the LLM will overestimate the impact of race, gender, etc. in its response. Similarly, when generating new clinical vignettes (e.g., for use in medical education), LLMs may over-index on demographic correlations (246), which would skew the knowledge of medical trainees if generated vignettes are widely used (248). It is likely that these issues can be mitigated through improved prompting and careful auditing of generated text. However, LLM stereotypes remain a key risk, both because they can be hard to detect, and because even small effect sizes can cause significant harm if the models are used at scale. Further work is necessary to better document such patterns, to properly inform users about them, and to develop mitigating strategies.

A second equity-related challenge is disparities in patient awareness of, and willingness to use, generative interfaces. Recent surveys show that awareness of LLMs positively correlates with formal education level and household income (249). Moreover, LLMs are more accessible to "tech-savvy" users (250), since using them requires a fast internet connection, intuition around how to best phrase prompts, etc. These factors raise the possibility that generative interfaces may create larger benefits for already-privileged patients who have fewer barriers to health access. (251, 252). To mitigate this risk, we need to study the factors underlying generative interface literacy (253) and ensure they become broadly accessible.

However, generative models also open important new health-equity-related *opportunities* (121, 254). (121) describes several such use cases for generative models: detecting human biases (e.g., from clinical notes); creating structured equity-relevant databases from unstructured text; and improving equity of access to health information. To identify such equity-related opportunities, it is imperative to focus on equity from the very beginning of a project, at the *problem selection* stage (226, 255).

3.7. Performing real-world evaluation

In order to reason about the real-world efficacy of generative models, we need fine-grained, real-world evaluations. Recent work has highlighted how evaluating clinical generative models on the basis of diagnosis alone overestimates their efficacy (70). Specifically, (70) highlight the importance of quantifying the extent to which generative models fail to adhere to treatment guidelines, or are sensitive to the order in which information is presented to them; measures of these types of behavior are important towards reasoning about the impact

of generative models in medicine.

The principles of fine-grained real-world evaluation apply to all use cases we highlight. For example, evaluations of generative models for medical question-answering must specify both (1) a set of prompts to test and (2) a set of criteria for a response to be deemed high-quality. For example, on (1), common medical LLM benchmarks use questions from exams (256, 257), clinical guidebooks (184, 187), or research papers (258). These evaluation sets may be systematically different from the real distribution of patient questions, both in medical content and linguistic characteristics (e.g., language, grammar, or dialect), which could affect performance (259–261). More recently, evaluations have included real patient questions from forums like MedlinePlus (262, 191), Reddit (`/r/AskDocs`; (263)), or cancer support groups on Facebook (264). These results may better reflect real-world patient questions, but future work should explicitly recruit patients from underserved groups, whose needs may differ from those represented by the dominant voices on online medical forums. On point (2), response quality is usually judged by physicians (191, 263, 265, 264), not patients. While physicians can best evaluate correctness, patients themselves may be better judges of perceived qualities like empathy (263) or understandability (187). Future work should also prioritize closing the gap between real-world usage and assumptions made during evaluation. For example, the standard approach to evaluation assumes that generative models have no capacity to collect additional information; a more realistic set-up would allow a generative model to pose follow-up questions (266). More broadly, a deeper understanding of the ways in which humans interact with generative interfaces will lead to a deeper understanding of generative model failures in the real world.

3.8. Clarifying accountability

The introduction of generative AI in healthcare raises important questions about who should be responsible for potential harms and regulation. Errors from generative models are inevitable, as they are with humans, but it remains unclear whether responsibility lies with the healthcare provider, the AI system developer, or the institution implementing the technology. Uncertainty surrounding liability is a key concern for healthcare providers who interact with generative interfaces (267), and regulation should be designed to reduce this uncertainty. Possible paths forward range from holding the "manufacturer" (i.e. the model developer) completely accountable for model errors to distributing responsibility across providers, hospitals, and model developers (268).

The question of accountability is further complicated by concerns about over-reliance. If healthcare professionals rely too heavily on generative AI tools to make accurate decisions, while model developers simultaneously rely on healthcare professionals to carefully vet those decisions, accountability may be lost. Healthcare professionals who rely too heavily on generative AI tools may not only find it difficult to make accurate decisions without them, but also be unable to detect errors when the generative AI tools are incorrect. Behavioral changes in response to the integration of AI tools are well-established (269, 270) and relate to the *autonomy* of decision-making, a key factor in determining liability (271). In the education space, researchers have found that students who have access to generative AI tools outperform a control group, but once the generative AI tools are removed, they perform worse (269). Similar studies have found that humans can also inherit biases from AI even when access to tools has been removed (270). These findings have important implications for the deployment of generative models in medicine, and suggest the importance of research
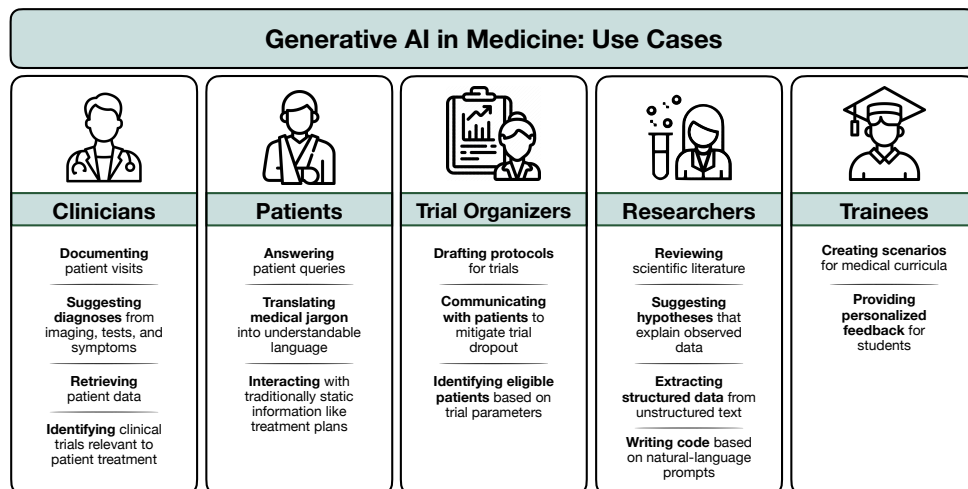
Generative AI in Medicine: Use Cases

Clinicians	Patients	Trial Organizers	Researchers	Trainees
Documenting patient visits	**Answering** patient queries	**Drafting protocols** for trials	**Reviewing** scientific literature	**Creating scenarios** for medical curricula
Suggesting diagnoses from imaging, tests, and symptoms	**Translating medical jargon** into understandable language	**Communicating with patients** to mitigate trial dropout	**Suggesting hypotheses** that explain observed data	**Providing personalized feedback** for students
Retrieving patient data	**Interacting** with traditionally static information like treatment plans	**Identifying eligible patients** based on trial parameters	**Extracting structured data** from unstructured text	
Identifying clinical trials relevant to patient treatment			**Writing code** based on natural-language prompts	

Figure 1

We highlight promising use cases of generative AI in medicine for five key constituent groups: clinicians, patients, trial organizers, researchers, and trainees.

and regulation that clarifies accountability in human-AI collaboration.

4. Conclusion

Those witnessing the explosion of interest in generative models in healthcare might justly feel both excitement and trepidation. On the one hand, increased model capabilities enable many use cases benefiting clinicians, patients, trial organizers, researchers, and trainees, with the potential to transform healthcare. But realizing this potential in high-stakes healthcare settings will require addressing numerous challenges — from centering equity, to protecting consent, to rigorously evaluating models — to bridge the gap between medical generative models in *theory* and in *practice*. The history of technical advances in medicine suggests that we will not be able to anticipate all the impacts of generative models — hospitals today are still adapting to the transition to EHRs, 15 years after their widespread introduction — and that humility is warranted. But, in the face of this uncertainty, the research directions we outline offer a roadmap for addressing generative AI's shortcomings, and realizing its potential, in order to provide better healthcare for all.

DISCLOSURE STATEMENT

Monica Agrawal is a co-founder of Layer Health and holds equity. Other authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.
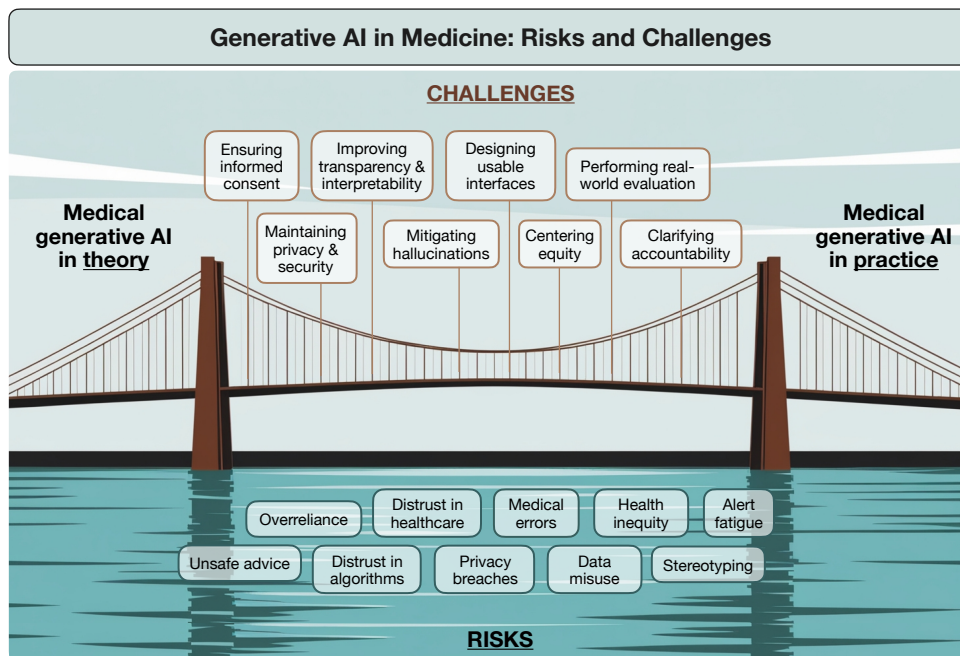
Generative AI in Medicine: Risks and Challenges

CHALLENGES

Ensuring informed consent | Improving transparency & interpretability | Designing usable interfaces | Performing real-world evaluation

Maintaining privacy & security | Mitigating hallucinations | Centering equity | Clarifying accountability

Medical generative AI in theory

Medical generative AI in practice

Overreliance | Distrust in healthcare | Medical errors | Health inequity | Alert fatigue

Unsafe advice | Distrust in algorithms | Privacy breaches | Data misuse | Stereotyping

RISKS

Figure 2

Bridging the gap between generative models in theory and practice will require addressing key challenges to mitigate risks and maximize benefits.

ACKNOWLEDGMENTS

LITERATURE CITED

1. Tai-Seale M, Baxter SL, Vaida F, Walker A, Sitapati AM, et al. 2024. AI-Generated Draft Replies Integrated Into Health Records and Physicians' Electronic Communication. *JAMA Network Open* 7(4):e246565–e246565

2. Zakka C, Cho J, Fahed G, Shad R, Moor M, et al. 2024. Almanac copilot: Towards autonomous electronic health record navigation. *arXiv preprint arXiv:2405.07896*

3. Garcia P, Ma SP, Shah S, Smith M, Jeong Y, et al. 2024. Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Network Open* 7(3):e243201–e243201

4. Kambhamettu H, Metaxa D, Johnson K, Head A. 2024. *Explainable Notes: Examining How to Unlock Meaning in Medical Notes with Interactivity and Artificial Intelligence*. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: Association for Computing Machinery

5. Mannhardt N, Bondi-Kelly E, Lam B, O'Connell C, Asiedu M, et al. 2024. Impact of large language model assistance on patients reading clinical notes: A mixed-methods study. *arXiv preprint arXiv:2401.09637*

6. Tierney AA, Gayre G, Hoberman B, Mattern B, Ballesca M, et al. 2024. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catalyst Innovations in Care Delivery* 5(3):CAT–23

7. Marshall IJ, Nye B, Kuiper J, Noel-Storr A, Marshall R, et al. 2020. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association* 27(12):1903–1912

8. Feldman J, Hochman KA, Guzman BV, Goodman A, Weisstuch J, Testa P. 2024. Scaling Note Quality Assessment Across an Academic Medical Center with AI and GPT-4. *NEJM Catalyst* 5(5):CAT.23.0283

9. Liu S, Wright AP, Mccoy AB, Huang SS, Genkins JZ, et al. 2024. Using large language model to guide patients to create efficient and comprehensive clinical care message. *Journal of the American Medical Informatics Association* :ocae142

10. Wang Z, Xiao C, Sun J. 2023. *AutoTrial: Prompting Language Models for Clinical Trial Design*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, ed. H Bouamor, J Pino, K Bali, pp. 12461–12472, pp. 12461–12472. Singapore: Association for Computational Linguistics

11. White RD, Peng T, Sripitak P, Johansen AR, Snyder M. 2023. *CliniDigest: A Case Study in Large Language Model Based Large-Scale Summarization of Clinical Trial Descriptions*. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*, pp. 396–402

12. Guo E, Gupta M, Deng J, Park YJ, Paget M, Naugler C. 2024. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *Journal of Medical Internet Research* 26:e48996

13. Ktena I, Wiles O, Albuquerque I, Rebuffi SA, Tanno R, et al. 2024. Generative models improve fairness of medical classifiers under distribution shifts. *Nature Medicine* 30(4):1166–1173

14. Bakkum MJ, Hartjes MG, Piët JD, Donker EM, Likic R, et al. 2024. Using artificial intelligence to create diverse and inclusive medical case vignettes for education. *British Journal of Clinical Pharmacology* 90(3):640–648

15. Dupont PE, Nelson BJ, Goldfarb M, Hannaford B, Menciassi A, et al. 2021. A decade retrospective of medical robotics research from 2010 to 2020. *Science robotics* 6(60):eabi8017

16. Kononenko I. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine* 23(1):89–109

17. Razzak MI, Imran M, Xu G. 2020. Big data analytics for preventive medicine. *Neural Computing and Applications* 32(9):4417–4451

18. Kruse CS, Smith B, Vanderlinden H, Nealand A. 2017. Security techniques for the electronic health records. *Journal of medical systems* 41:1–9

19. Fernández-Alemán JL, Señor IC, Lozoya PÁO, Toval A. 2013. Security and privacy in electronic health records: A systematic literature review. *Journal of biomedical informatics* 46(3):541–562

20. Campanella P, Lovato E, Marone C, Fallacara L, Mancuso A, et al. 2016. The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. *The European Journal of Public Health* 26(1):60–64

21. Graber ML, Siegal D, Riah H, Johnston D, Kenyon K. 2019. Electronic health record–related events in medical malpractice claims. *Journal of patient safety* 15(2):77–85

22. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, et al. 2017. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC medical informatics and decision making* 17:1–9

23. Rosenberg N. 1982. Learning by using. *Inside the black box: Technology and economics* :120–

140

24. Bishop CM, Nasrabadi NM. 2006. *Pattern recognition and machine learning*, vol. 4. Springer

25. Kaplan J, McCandlish S, Henighan T, Brown TB, Chess B, et al. 2020. Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs, stat]* ArXiv: 2001.08361

26. Movva R, Balachandar S, Peng K, Agostini G, Garg N, Pierson E. 2024. *Topics, Authors, and Institutions in Large Language Model Research: Trends from 17K arXiv Papers*. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1223–1243

27. Abbaspourazad S, Elachqar O, Miller AC, Emrani S, Nallasamy U, Shapiro I. 2023. Large-scale Training of Foundation Models for Wearable Biosignals

28. Beaini D, Huang S, Cunha JA, Li Z, Moisescu-Pareja G, et al. 2023. Towards Foundational Models for Molecular Learning on Large-Scale Multi-Task Datasets

29. McKeen K, Oliva L, Masood S, Toma A, Rubin B, Wang B. 2024. ECG-FM: An Open Electrocardiogram Foundation Model

30. Bond-Taylor S, Leach A, Long Y, Willcocks CG. 2021. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models

31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. 2017. Attention is all you need. *Advances in neural information processing systems* 30

32. Jurafsky D. 2000. Speech and language processing

33. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, et al. 2022. Training language models to follow instructions with human feedback. ArXiv:2203.02155 [cs]

34. Bolton E, Venigalla A, Yasunaga M, Hall D, Xiong B, et al. 2024. BioMedLM: A 2.7B Parameter Language Model Trained On Biomedical Text. ArXiv:2403.18421 [cs]

35. Chen Z, Cano AH, Romanou A, Bonnet A, Matoba K, et al. 2023. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models. ArXiv:2311.16079 [cs]

36. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, et al. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. ArXiv:2305.09617 [cs]

37. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences* 11(14):6421Number: 14 Publisher: Multidisciplinary Digital Publishing Institute

38. Fleming SL, Lozano A, Haberkorn WJ, Jindal JA, Reis EP, et al. 2023. MedAlign: A Clinician-Generated Dataset for Instruction Following with Electronic Medical Records. ArXiv:2308.14089 [cs]

39. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine* 4(1):1–13Publisher: Nature Publishing Group

40. Hill BL, Emami M, Nori VS, Cordova-Palomera A, Tillman RE, Halperin E. 2023. *CHIRon: A Generative Foundation Model for Structured Sequential Medical Data*

41. Ferruz N, Schmidt S, Höcker B. 2022. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications* 13(1):4348

42. Rives A, Meier J, Sercu T, Goyal S, Lin Z, et al. 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* 118(15):e2016239118Publisher: Proceedings of the National Academy of Sciences

43. Nguyen E, Poli M, Durrant MG, Thomas AW, Kang B, et al. 2024. Sequence modeling and design from molecular to genome scale with Evo. Pages: 2024.02.27.582234 Section: New Results

44. Yang L, Zhang Z, Song Y, Hong S, Xu R, et al. 2023. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.* 56(4):105:1–105:39

45. Dhariwal P, Nichol A. 2021. Diffusion Models Beat GANs on Image Synthesis. ArXiv:2105.05233 [cs, stat]

46. Kazerouni A, Aghdam EK, Heidari M, Azad R, Fayyaz M, et al. 2023. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis* 88:102846

47. Sun S, Goldgof G, Butte A, Alaa AM. 2023. Aligning Synthetic Medical Images with Clinical Knowledge using Human Feedback. *Advances in Neural Information Processing Systems* 36:13408–13428

48. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, et al. 2021. *Learning Transferable Visual Models From Natural Language Supervision*. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR. ISSN: 2640-3498

49. Chambon PJM, Bluethgen C, Langlotz C, Chaudhari A. 2022. *Adapting Pretrained Vision-Language Foundational Models to Medical Imaging Domains*

50. Li J, Li D, Savarese S, Hoi S. 2023. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742. PMLR. ISSN: 2640-3498

51. Thawakar OC, Shaker AM, Mullappilly SS, Cholakkal H, Anwer RM, et al. 2024. *XrayGPT: Chest Radiographs Summarization using Large Medical Vision-Language Models*. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, ed. D Demner-Fushman, S Ananiadou, M Miwa, K Roberts, J Tsujii, pp. 440–448, pp. 440–448. Bangkok, Thailand: Association for Computational Linguistics

52. Bazi Y, Rahhal MMA, Bashmal L, Zuair M. 2023. Vision–Language Model for Visual Question Answering in Medical Imagery. *Bioengineering* 10(3):380Number: 3 Publisher: Multidisciplinary Digital Publishing Institute

53. Li C, Wong C, Zhang S, Usuyama N, Liu H, et al. 2023. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *Advances in Neural Information Processing Systems* 36:28541–28564

54. Aiken LH, Lasater KB, Sloane DM, Pogue CA, Fitzpatrick Rosenbaum KE, et al. 2023. Physician and Nurse Well-Being and Preferred Interventions to Address Burnout in Hospital Practice: Factors Associated With Turnover, Outcomes, and Patient Safety. *JAMA Health Forum* 4(7):e231809

55. Saag HS, Shah K, Jones SA, Testa PA, Horwitz LI. 2019. Pajama time: working after work in the electronic health record. *Journal of general internal medicine* 34:1695–1696

56. Heer J. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116(6):1844–1850

57. Mamykina L, Vawdrey DK, Stetson PD, Zheng K, Hripcsak G. 2012. Clinical documentation: composition or synthesis? *Journal of the American Medical Informatics Association* 19(6):1025–1031

58. Jiang S, Shen S, Agrawal M, Lam B, Kurtzman N, et al. 2023. *Conceptualizing machine learning for dynamic information retrieval of electronic health record notes*. In *Machine Learning for Healthcare Conference*, pp. 343–359. PMLR

59. Young CC, Enichen E, Rivera C, Auger CA, Grant N, et al. 2024. Diagnostic Accuracy of a Custom Large Language Model on Rare Pediatric Disease Case Reports. *American Journal of Medical Genetics Part A* n/a(n/a):e63878

60. Kanjee Z, Crowe B, Rodman A. 2023. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. *JAMA*

61. Levine DM, Tuwani R, Kompa B, Varma A, Finlayson SG, et al. 2023. The Diagnostic and Triage Accuracy of the GPT-3 Artificial Intelligence Model

62. Ríos-Hoyo A, Shan NL, Li A, Pearson AT, Pusztai L, Howard FM. 2024. Evaluation of large language models as a diagnostic aid for complex medical cases. *Frontiers in Medicine* 11

63. Olmo Jd, Logroño J, Mascías C, Martínez M, Isla J. 2024. Assessing DxGPT: Diagnosing Rare Diseases with Various Large Language Models

64. Kanjee Z, Crowe B, Rodman A. 2023. Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge. *JAMA* 330(1):78–80

65. Zhou J, He X, Sun L, Xu J, Chen X, et al. 2024. Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4. *Nature Communications* 15(1):5649

66. Thawkar O, Shaker A, Mullappilly SS, Cholakkal H, Anwer RM, et al. 2023. XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models

67. Moor M, Huang Q, Wu S, Yasunaga M, Zakka C, et al. 2023. Med-Flamingo: a Multimodal Medical Few-shot Learner

68. Lin Z, Zhang D, Tao Q, Shi D, Haffari G, et al. 2023. Medical visual question answering: A survey. *Artificial Intelligence in Medicine* 143:102611

69. Reese JT, Danis D, Caufield JH, Groza T, Casiraghi E, et al. 2024. On the limitations of large language models in clinical diagnosis. *medRxiv* :2023.07.13.23292613

70. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine* 30(9):2613–2622

71. Ahmed A, Chandra S, Herasevich V, Gajic O, Pickering BW. 2011. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. *Critical care medicine* 39(7):1626–1634

72. Murray L, Gopinath D, Agrawal M, Horng S, Sontag D, Karger DR. 2021. *Medknowts: unified documentation and information retrieval for electronic health records*. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pp. 1169–1183

73. Zheng K, Padman R, Johnson MP, Diamond HS. 2009. An interface-driven analysis of user interactions with an electronic health records system. *Journal of the American Medical Informatics Association* 16(2):228–237

74. Sackett DL, Rosenberg WMC. 1995. On the need for evidence-based medicine. *Journal of Public Health* 17(3):330–334

75. Bastian H, Glasziou P, Chalmers I. 2010. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine* 7(9):e1000326

76. Zakka C, Chaurasia A, Shad R, Dalal AR, Kim JL, et al. 2023. Almanac: Retrieval-Augmented Language Models for Clinical Medicine. ArXiv:2303.01229 [cs]

77. Lee K, Paek H, Huang LC, Hilton CB, Datta S, et al. 2024. SEETrials: Leveraging Large Language Models for Safety and Efficacy Extraction in Oncology Clinical Trials

78. Chewning B, Bylund CL, Shah B, Arora NK, Gueguen JA, Makoul G. 2012. Patient preferences for shared decisions: a systematic review. *Patient education and counseling* 86(1):9–18

79. Adler RF, Morales P, Sotelo J, Magasi S. 2022. Developing an mhealth app for empowering cancer survivors with disabilities: Co-design study. *JMIR Formative Research* 6(7):e37706

80. Noack EM, Schulze J, Müller F. 2021. Designing an app to overcome language barriers in the delivery of emergency medical services: participatory development process. *JMIR mHealth and uHealth* 9(4):e21586

81. Danieli M, Ciulli T, Mousavi SM, Riccardi G. 2021. A conversational artificial intelligence agent for a mental health care app: Evaluation study of its participatory design. *JMIR Formative Research* 5(12):e30053

82. Martin-Hammond A, Vemireddy S, Rao K, et al. 2019. Exploring older adults' beliefs about the use of intelligent assistants for consumer health information management: A participatory design study. *JMIR aging* 2(2):e15381

83. Amante DJ, Hogan TP, Pagoto SL, English TM, Lapane KL. 2015. Access to care and use of the internet to search for health information: results from the us national health interview survey. *Journal of medical Internet research* 17(4):e106

84. Thapa DK, Visentin DC, Kornhaber R, West S, Cleary M. 2021. The influence of online health information on health decisions: A systematic review. *Patient education and counseling* 104(4):770–784

85. Vanessa Choy, Sara Martin, Ashley Lumpkin. 2024. Can we rely on generative AI for healthcare information? | Ipsos

86. Alex Montero, Grace Sparks, Marley Presiado, Liz Hamel. 2024. KFF Health Misinformation Tracking Poll: Health and Election Issues on TikTok | KFF

87. Hersh W. 2024. Search still matters: information retrieval in the era of generative ai. *Journal of the American Medical Informatics Association* :ocae014

88. Grossman LV, Feiner SK, Mitchell EG, Creber RMM. 2018. Leveraging patient-reported outcomes using data visualization. *Applied clinical informatics* 9(03):565–575

89. Zhao JY, Song B, Anand E, Schwartz D, Panesar M, et al. 2017. *Barriers, facilitators, and solutions to optimal patient portal and personal health record use: a systematic review of the literature.* In *AMIA annual symposium proceedings*, vol. 2017, pp. 1913. American Medical Informatics Association

90. Grossman LV, Choi SW, Collins S, Dykes PC, O'Leary KJ, et al. 2018. Implementation of acute care patient portals: recommendations on utility and use from six early adopters. *Journal of the American Medical Informatics Association* 25(4):370–379

91. Grossman LV, Masterson Creber RM, Benda NC, Wright D, Vawdrey DK, Ancker JS. 2019. Interventions to increase patient portal use in vulnerable populations: a systematic review. *Journal of the American Medical Informatics Association* 26(8-9):855–870

92. Warren LR, Harrison M, Arora S, Darzi A. 2019. Working with patients and the public to design an electronic health record interface: a qualitative mixed-methods study. *BMC medical informatics and decision making* 19:1–8

93. Kambhamettu H, Metaxa D, Johnson K, Head A. 2024. *Explainable Notes: Examining How to Unlock Meaning in Medical Notes with Interactivity and Artificial Intelligence.* In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–19

94. Luo L, Vairavamurthy J, Zhang X, Kumar A, Ter-Oganesyan RR, et al. 2024. ReXplain: Translating Radiology into Patient-Friendly Video Reports

95. Basu C, Vasu R, Yasunaga M, Yang Q. 2023. *Med-EASi: finely annotated dataset and models for controllable simplification of medical texts.* In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press

96. Mirza FN, Tang OY, Connolly ID, Abdulrazeq HA, Lim RK, et al. 2024. Using chatgpt to facilitate truly informed medical consent. *NEJM AI* 1(2):AIcs2300145

97. Shadbolt C, Naufal E, Bunzli S, Price V, Rele S, et al. 2023. Analysis of Rates of Completion, Delays, and Participant Recruitment in Randomized Clinical Trials in Surgery. *JAMA Network Open* 6(1):e2250996

98. Ross JS, Mocanu M, Lampropulos JF, Tse T, Krumholz HM. 2013. Time to Publication Among Completed Clinical Trials. *JAMA Internal Medicine* 173(9):825–828

99. Zarin DA, Tse T, Williams RJ, Rajakannan T. 2017. Update on Trial Registration 11 Years after the ICMJE Policy Was Established. *New England Journal of Medicine* 376(4):383–391

100. Getz KA, Stergiopoulos S, Short M, Surgeon L, Krauss R, et al. 2016. The Impact of Protocol Amendments on Clinical Trial Performance and Cost. *Therapeutic Innovation & Regulatory Science* 50(4):436–441

101. Fogel DB. 2018. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary Clinical Trials Communications* 11:156–164

102. Ghim JL, Ahn S. 2023. Transforming clinical trials: the emerging roles of large language models. *Translational and Clinical Pharmacology* 31(3):131–138

103. Park J, Fang Y, Ta C, Zhang G, Idnay B, et al. 2024. Criteria2Query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. *Journal of Biomedical Informatics* 154:104649

104. Lai H, Ge L, Sun M, Pan B, Huang J, et al. 2024. Assessing the Risk of Bias in Randomized

Clinical Trials With Large Language Models. *JAMA Network Open* 7(5):e2412687

105. Datta S, Lee K, Paek H, Manion FJ, Ofoegbu N, et al. 2024. AutoCriteria: a generalizable clinical trial eligibility criteria extraction system powered by large language models. *Journal of the American Medical Informatics Association* 31(2):375–385

106. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, et al. 2019. Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association* 26(4):294–305

107. Hamer DMd, Schoor P, Polak TB, Kapitan D. 2023. Improving Patient Pre-screening for Clinical Trials: Assisting Physicians with Large Language Models. ArXiv:2304.07396 [cs]

108. Wornow M, Lozano A, Dash D, Jindal J, Mahaffey KW, Shah NH. 2024. Zero-Shot Clinical Trial Patient Matching with LLMs

109. Jin Q, Wang Z, Floudas CS, Chen F, Gong C, et al. 2024. Matching Patients to Clinical Trials with Large Language Models. *ArXiv* :arXiv:2307.15051v4

110. Beattie J, Neufeld S, Yang D, Chukwuma C, Gul A, et al. 2024. Utilizing Large Language Models for Enhanced Clinical Trial Matching: A Study on Automation in Patient Screening. *Cureus* 16(5):e60044

111. McCann S, Campbell M, Entwistle V. 2013. Recruitment to clinical trials: a meta-ethnographic synthesis of studies of reasons for participation. *Journal of Health Services Research & Policy* 18(4):233–241

112. Skea ZC, Newlands R, Gillies K. 2019. Exploring non-retention in clinical trials: a meta-ethnographic synthesis of studies reporting participant reasons for drop out. *BMJ Open* 9(6):e021959

113. Goodson N, Wicks P, Morgan J, Hashem L, Callinan S, Reites J. 2022. Opportunities and counterintuitive challenges for decentralized clinical trials to broaden participant inclusion. *NPJ Digital Medicine* 5:58

114. Thomas KA, Kidziński L. 2022. Artificial intelligence can improve patients' experience in decentralized clinical trials. *Nature Medicine* 28(12):2462–2463

115. Zhou Q, Ratcliffe SJ, Grady C, Wang T, Mao JJ, Ulrich CM. 2019. Cancer Clinical Trial Patient-Participants' Perceptions about Provider Communication and Dropout Intentions. *AJOB empirical bioethics* 10(3):190–200

116. Dennstädt F, Zink J, Putora PM, Hastings J, Cihoric N. 2024. Title and abstract screening for literature reviews using large language models: an exploratory study in the biomedical domain. *Systematic Reviews* 13(1):158

117. Chen RJ, Lu MY, Chen TY, Williamson DFK, Mahmood F. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* 5(6):493–497

118. Khosravi B, Li F, Dapamede T, Rouzrokh P, Gamble CU, et al. 2024. Synthetically enhanced: unveiling synthetic data's potential in medical imaging research. *eBioMedicine* 104:105174

119. Das HP, Tran R, Singh J, Yue X, Tison G, et al. 2021. Conditional Synthetic Data Generation for Robust Machine Learning Applications with Limited Pandemic Data

120. Ive J, Viani N, Kam J, Yin L, Verma S, et al. 2020. Generation and evaluation of artificial mental health records for Natural Language Processing. *npj Digital Medicine* 3(1):1–9

121. Pierson E, Shanmugam D, Movva R, Kleinberg J, Agrawal M, et al. 2024. Use large language models to promote health equity. *arXiv preprint arXiv:2312.14804*

122. Sun Y, Zhu C, Zheng S, Zhang K, Sun L, et al. 2024. PathAsst: A Generative Foundation AI Assistant Towards Artificial General Intelligence of Pathology

123. Lu MY, Chen B, Williamson DFK, Chen RJ, Zhao M, et al. 2024. A Multimodal Generative AI Copilot for Human Pathology. *Nature* :1–3

124. Huang Z, Bianchi F, Yuksekgonul M, Montine TJ, Zou J. 2023. A visual-language foundation model for pathology image analysis using medical Twitter. *Nature Medicine* 29(9):2307–2316

125. Wang H, Fu T, Du Y, Gao W, Huang K, et al. 2023. Scientific discovery in the age of artificial intelligence. *Nature* 620(7972):47–60

126. Wong HE, Rakic M, Guttag J, Dalca AV. 2024. ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Biomedical Image

127. Zhou Y, Liu H, Srivastava T, Mei H, Tan C. 2024. Hypothesis Generation with Large Language Models

128. Zhong R, Zhang P, Li S, Ahn J, Klein D, Steinhardt J. 2023. Goal Driven Discovery of Distributional Differences via Language Descriptions

129. Pham CM, Hoyle A, Sun S, Resnik P, Iyyer M. 2024. TopicGPT: A Prompt-based Topic Modeling Framework

130. Kamienny PA, d'Ascoli S, Lample G, Charton F. 2022. End-to-end symbolic regression with transformers. *Advances in Neural Information Processing Systems* 35:10269–10281

131. Tayebi Arasteh S, Han T, Lotfinia M, Kuhl C, Kather JN, et al. 2024. Large language models streamline automated machine learning for clinical studies. *Nature Communications* 15(1):1603

132. Mozannar H, Chen V, Alsobay M, Das S, Zhao S, et al. 2024. The RealHumanEval: Evaluating Large Language Models' Abilities to Support Programmers

133. Biri SK, Kumar S, Panigrahi M, Mondal S, Behera JK, Mondal H. 2024. Assessing the Utilization of Large Language Models in Medical Education: Insights From Undergraduate Medical Students. *Cureus* 15(10):e47468

134. Grigorian A, Shipley J, Nahmias J, Nguyen N, Schwed AC, et al. 2023. Implications of Using Chatbots for Future Surgical Education. *JAMA Surgery* 158(11):1220–1222

135. Lee CR, Gilliland KO, Beck Dallaghan GL, Tolleson-Rinehart S. 2022. Race, ethnicity, and gender representation in clinical case vignettes: a 20-year comparison between two institutions. *BMC Medical Education* 22(1):585

136. Benoit JR. 2023. Chatgpt for clinical vignette generation, revision, and evaluation. *medRxiv* :2023–02

137. Tejani AS, Elhalawani H, Moy L, Kohli M, Kahn CE. 2023. Artificial Intelligence and Radiology Education. *Radiology: Artificial Intelligence* 5(1):e220084

138. Holderried F, Stegemann-Philipps C, Herschbach L, Moldt JA, Nevins A, et al. 2024. A Generative Pretrained Transformer (GPT)-Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study. *JMIR medical education* 10:e53961

139. Dai W, Lin J, Jin H, Li T, Tsai YS, et al. 2023. *Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT*. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pp. 323–325

140. Belmar F, Gaete MI, Escalona G, Carnier M, Durán V, et al. 2023. Artificial intelligence in laparoscopic simulation: a promising future for large-scale automated evaluations. *Surgical Endoscopy* 37(6):4942–4946

141. Zhao Y, Wang Y, Zhang J, Liu X, Li Y, et al. 2022. Surgical GAN: Towards real-time path planning for passive flexible tools in endovascular surgeries. *Neurocomputing* 500:567–580

142. Kirby MD. 1983. Informed consent: what does it mean? *Journal of medical ethics* 9(2):69–75

143. Riddick FA. 2003. The code of medical ethics of the american medical association

144. Del Carmen MG, Joffe S. 2005. Informed consent for medical treatment and research: a review. *The oncologist* 10(8):636–641

145. Pierson L, Pierson E. 2022. Patients cannot consent to care unless they know how much it costs

146. Astromskė K, Peičius E, Astromskis P. 2021. Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. *AI & SOCIETY* 36:509–520

147. Wilcox L, Brewer R, Diaz F. 2023. Ai consent futures: A case study on voice data collection with clinicians. *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW2):1–30

148. Garcia Valencia OA, Suppadungsuk S, Thongprayoon C, Miao J, Tangpanithandee S, et al. 2023. Ethical implications of chatbot utilization in nephrology. *Journal of Personalized Medicine* 13(9):1363

149. Decker H, Trang K, Ramirez J, Colley A, Pierce L, et al. 2023. Large language model- based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Network Open* 6(10):e2336997–e2336997

150. Burks AC, Keim-Malpass J. 2019. Health literacy and informed consent for clinical trials: a systematic review and implications for nurses. *Nursing: Research and Reviews* :31–40

151. Simon C, Zyzanski SJ, Eder M, Raiz P, Kodish ED, Siminoff LA. 2003. Groups potentially at risk for making poorly informed decisions about entry into clinical trials for childhood cancer. *Journal of Clinical Oncology* 21(11):2173–2178

152. Raimann FJ, Neef V, Hennighausen MC, Zacharowski K, Flinspach AN. 2024. Evaluation of ai chatbots for the creation of patient-informed consent sheets. *Machine Learning and Knowledge Extraction* 6(2):1145–1153

153. Chen Y, Esmaeilzadeh P. 2024. Generative ai in medical practice: in-depth exploration of privacy and security challenges. *Journal of Medical Internet Research* 26:e53008

154. Bai X, Wang H, Ma L, Xu Y, Gan J, et al. 2021. Advancing covid-19 diagnosis with privacy-preserving collaboration in artificial intelligence. *Nature Machine Intelligence* 3(12):1081–1089

155. Ali M, Naeem F, Tariq M, Kaddoum G. 2022. Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey. *IEEE journal of biomedical and health informatics* 27(2):778–789

156. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. 2021. Federated learning for healthcare informatics. *Journal of healthcare informatics research* 5:1–19

157. Geiping J, Bauermeister H, Dröge H, Moeller M. 2020. Inverting Gradients – How easy is it to break privacy in federated learning?

158. So J, Ali RE, Guler B, Jiao J, Avestimehr S. 2023. Securing Secure Aggregation: Mitigating Multi-Round Privacy Leakage in Federated Learning

159. Huang Y, Gupta S, Song Z, Li K, Arora S. 2021. Evaluating Gradient Inversion Attacks and Defenses in Federated Learning

160. Mullainathan S, Obermeyer Z. 2022. Solving medicine's data bottleneck: Nightingale open science. *Nature Medicine* 28(5):897–899

161. Johnson AE, Bulgarelli L, Shen L, Gayles A, Shammout A, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data* 10(1):1

162. Barrett C, Boyd B, Bursztein E, Carlini N, Chen B, et al. 2023. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security* 6(1):1–52

163. El-Mhamdi EM, Farhadkhani S, Guerraoui R, Gupta N, Hoang LN, et al. 2022. On the impossible safety of large ai models. *arXiv preprint arXiv:2209.15259*

164. Carlini N, Ippolito D, Jagielski M, Lee K, Tramer F, Zhang C. 2022. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*

165. Huang J, Shao H, Chang KCC. 2022. Are large pre-trained language models leaking your personal information? *arXiv preprint arXiv:2205.12628*

166. Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, et al. 2021. *Extracting training data from large language models*. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650

167. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. 2018. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks

168. Ghosheh GO, Li J, Zhu T. 2024. A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Computing Surveys* 56(6):1–34

169. Loong B, Zaslavsky AM, He Y, Harrington DP. 2013. Disclosure Control using Partially Synthetic Data for Large-Scale Health Surveys, with Applications to CanCORS. *Statistics in medicine* 32(24):4139–4161

170. Bommasani R, Klyman K, Longpre S, Kapoor S, Maslej N, et al. 2023. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*

171. Bommasani R, et al. 2024. The foundation model transparency index v1.1 may 2024. *Stanford*

CRFM

172. Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, et al. 2019. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology* 155(10):1135–1141

173. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. 2020. *Hidden stratification causes clinically meaningful failures in machine learning for medical imaging.* In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159

174. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. 2018. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*

175. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. 2018. *Explaining explanations: An overview of interpretability of machine learning.* In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89. IEEE

176. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. 2020. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10(5):e1379

177. Ghassemi M, Oakden-Rayner L, Beam AL. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3(11):e745–e750

178. Bilodeau B, Jaques N, Koh PW, Kim B. 2024. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences* 121(2):e2304406120

179. Zhao H, Yang F, Lakkaraju H, Du M. 2024. Opening the black box of large language models: Two views on holistic interpretability. *arXiv preprint arXiv:2402.10688*

180. Singh C, Inala JP, Galley M, Caruana R, Gao J. 2024. Rethinking interpretability in the era of large language models. *arXiv preprint arXiv:2402.01761*

181. Agarwal C, Tanneru SH, Lakkaraju H. 2024. Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. *arXiv preprint arXiv:2402.04614*

182. U N, M K, J K. 2023. Vision-Language Transformer for Interpretable Pathology Visual Question Answering. *IEEE journal of biomedical and health informatics* 27(4)

183. Kim C, Gadgil SU, DeGrave AJ, Omiye JA, Cai ZR, et al. 2024. Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature Medicine* :1–12

184. Chen S, Kann BH, Foote MB, Aerts HJWL, Savova GK, et al. 2023. Use of Artificial Intelligence Chatbots for Cancer Treatment Information. *JAMA Oncology*

185. Ji Z, Lee N, Frieske R, Yu T, Su D, et al. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* 55(12):248:1–248:38

186. Lee P, Bubeck S, Petro J. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine* 388(13):1233–1239

187. Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. 2023. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. *JAMA Oncology*

188. Shuster K, Poff S, Chen M, Kiela D, Weston J. 2021. Retrieval Augmentation Reduces Hallucination in Conversation

189. Agrawal M, Hegselmann S, Lang H, Kim Y, Sontag D. 2022. *Large language models are few-shot clinical information extractors.* In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1998–2022

190. Gero Z, Singh C, Cheng H, Naumann T, Galley M, et al. 2023. Self-Verification Improves Few-Shot Clinical Information Extraction. ArXiv:2306.00024 [cs]

191. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, et al. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. ArXiv:2305.09617 [cs]

192. Meskó B, Topol EJ. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *npj Digital Medicine* 6(1):1–6Number: 1 Publisher: Nature Publishing Group

193. Jakob Nielsen. 2023. AI: First New UI Paradigm in 60 Years. *Nielsen Norman Group*

194. Sai S, Gaur A, Sai R, Chamola V, Guizani M, Rodrigues JJ. 2024. Generative AI for transformative healthcare: A comprehensive study of emerging models, applications, case studies and limitations. *IEEE Access*

195. Mulia AP, Piri PR, Tho C. 2023. Usability analysis of text generation by chatgpt openai using system usability scale method. *Procedia Computer Science* 227:381–388

196. Giunti G, Doherty CP. 2024. Cocreating an automated mhealth apps systematic review process with generative ai: Design science research approach. *JMIR Medical Education* 10:e48949

197. Tankelevitch L, Kewenig V, Simkute A, Scott AE, Sarkar A, et al. 2024. *The metacognitive demands and opportunities of generative AI*. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–24

198. Dang H, Mecke L, Lehmann F, Goller S, Buschek D. 2022. How to prompt? Opportunities and challenges of zero-and few-shot learning for human-AI interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390*

199. Sun J, Liao QV, Muller M, Agarwal M, Houde S, et al. 2022. *Investigating explainability of generative AI for code through scenario-based design*. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pp. 212–228

200. Subramonyam H, Pea R, Pondoc C, Agrawala M, Seifert C. 2024. *Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs*. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–19

201. Zamfirescu-Pereira J, Wong RY, Hartmann B, Yang Q. 2023. *Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts*. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21

202. Abbasian M, Khatibi E, Azimi I, Oniani D, Shakeri Hossein Abad Z, et al. 2024. Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative ai. *NPJ Digital Medicine* 7(1):82

203. Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7(CSCW1):1–38

204. Kostick-Quenet KM, Gerke S. 2022. AI in the hands of imperfect users. *npj Digital Medicine* 5(1):197

205. Wysocki O, Davies JK, Vigo M, Armstrong AC, Landers D, et al. 2023. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence* 316:103839

206. Jacobs M, Pradier MF, McCoy Jr TH, Perlis RH, Doshi-Velez F, Gajos KZ. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11(1):108

207. Weisz JD, He J, Muller M, Hoefer G, Miles R, Geyer W. 2024. *Design Principles for Generative AI Applications*. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–22

208. Schaefer KE, Chen JY, Szalma JL, Hancock PA. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58(3):377–400

209. Mertz L. 2015. From annoying to appreciated: Turning clinical decision support systems into a medical professional's best friend. *IEEE pulse* 6(5):4–9

210. Greenes RA, Bates DW, Kawamoto K, Middleton B, Osheroff J, Shahar Y. 2018. Clinical decision support models and frameworks: seeking to address research issues underlying implementation successes and failures. *Journal of biomedical informatics* 78:134–143

211. Wright A, Hickman TTT, McEvoy D, Aaron S, Ai A, et al. 2016. Analysis of clinical decision support system malfunctions: a case series and survey. *Journal of the American Medical Informatics Association* 23(6):1068–1076

212. Khan S, Richardson S, Liu A, Mechery V, McCullagh L, et al. 2019. Improving provider

adoption with adaptive clinical decision support surveillance: an observational study. *JMIR human factors* 6(1):e10245

213. Mann D, Hess R, McGinn T, Richardson S, Jones S, et al. 2020. Impact of clinical decision support on antibiotic prescribing for acute respiratory infections: a cluster randomized implementation trial. *Journal of General Internal Medicine* 35:788–795

214. Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, et al. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4(1):31

215. Gaube S, Suresh H, Raue M, Lermer E, Koch TK, et al. 2023. Non-task expert physicians benefit from correct explainable ai advice when reviewing x-rays. *Scientific reports* 13(1):1383

216. Jacobs M, He J, F. Pradier M, Lam B, Ahn AC, et al. 2021. *Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens*. In *Proceedings of the 2021 chi conference on human factors in computing systems*, pp. 1–14

217. Henry KE, Adams R, Parent C, Soleimani H, Sridharan A, et al. 2022. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nature medicine* 28(7):1447–1454

218. Cabrera AA, Perer A, Hong JI. 2023. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proc. ACM Hum.-Comput. Interact.* 7(CSCW1):136:1–136:21

219. Bansal G, Nushi B, Kamar E, Lasecki WS, Weld DS, Horvitz E. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7:2–11

220. 2024. Gen AI Saves Nurses Time by Drafting Responses to Patient Messages — epicshare.org. https://www.epicshare.org/share-and-learn/mayo-ai-message-responses

221. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453

222. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, et al. 2022. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health* 4(6):e406–e414

223. Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, et al. 2022. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances* 8(31):eabq6147

224. Gervasi SS, Chen IY, Smith-McLallen A, Sontag D, Obermeyer Z, et al. 2022. The potential for bias in machine learning and opportunities for health insurers to address it. *Health Affairs* 41(2):212–218

225. Zink A, Obermeyer Z, Pierson E. 2024. Race adjustments in clinical algorithms can help correct for racial disparities in data quality. *PNAS*

226. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. 2021. Ethical machine learning in healthcare. *Annual review of biomedical data science* 4:123–144

227. Pierson E. 2020. Assessing racial inequality in covid-19 testing with bayesian threshold tests. *Extended abstract, NeurIPS ML4H*

228. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, et al. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine* 25(9):1337–1340

229. Pierson E. 2024. Accuracy and equity in clinical risk prediction. *The New England Journal of Medicine* 390(2):100–102

230. Seyyed-Kalantari L, Zhang H, McDermott MB, Chen IY, Ghassemi M. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine* 27(12):2176–2182

231. Movva R, Shanmugam D, Hou K, Pathak P, Guttag J, et al. 2023. *Coarse race data conceals disparities in clinical risk score performance*. In *Machine Learning for Healthcare Conference*, pp. 443–472. PMLR

232. Zou J, Gichoya JW, Ho DE, Obermeyer Z. 2023. Implications of predicting race variables from medical images. *Science* 381(6654):149–150

233. Balachandar S, Garg N, Pierson E. 2024. Domain constraints improve risk prediction when outcome data is missing. *ICLR*

234. Ferryman K, Mackintosh M, Ghassemi M. 2023. Considering biased data as informative artifacts in AI-assisted health care. *New England Journal of Medicine* 389(9):833–838

235. Shanmugam D, Hou K, Pierson E. 2024. Quantifying disparities in intimate partner violence: a machine learning method to correct for underreporting. *npj Women's Health* 2(1):15

236. Zink A, Rose S. 2020. Fair regression for health care spending. *Biometrics* 76(3):973–982

237. Vyas DA, Eisenstein LG, Jones DS. 2020. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms

238. Mullainathan S, Obermeyer Z. 2021. *On the inequity of predicting A while hoping for B.* In *AEA Papers and Proceedings*, vol. 111, pp. 37–42. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203

239. Pierson E, Cutler DM, Leskovec J, Mullainathan S, Obermeyer Z. 2021. An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine* 27(1):136–140

240. Obermeyer Z, Nissan R, Stern M, Eaneff S, Bembeneck EJ, Mullainathan S. 2021. Algorithmic bias playbook. *Center for Applied AI at Chicago Booth*

241. Diao JA, He Y, Khazanchi R, Tiako MN, Witonsky JI, et al. 2024. Implications of race adjustment in lung-function equations. *The New England journal of medicine* 390(22):2083

242. Diao JA, Shi I, Murthy VL, Buckley TA, Patel CJ, et al. 2024. Projected changes in statin and antihypertensive therapy eligibility with the aha prevent cardiovascular risk equations. *JAMA* 332(12):989–1000

243. Pfohl SR, Cole-Lewis H, Sayres R, Neal D, Asiedu M, et al. 2024. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine* :1–11

244. Hofmann V, Kalluri PR, Jurafsky D, King S. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature* 633(8028):147–154

245. Hoffman KM, Trawalter S, Axt JR, Oliver MN. 2016. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences of the United States of America* 113(16):4296–4301

246. Zack T, Lehman E, Suzgun M, Rodriguez JA, Celi LA, et al. 2024. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health* 6(1):e12–e22

247. Omiye JA, Lester J, Spichak S, Rotemberg V, Daneshjou R. 2023. Beyond the hype: large language models propagate race-based medicine. *medRxiv* :2023–07

248. Ripp K, Braun L. 2017. Race/Ethnicity in Medical Education: An Analysis of a Question Bank for Step 1 of the United States Medical Licensing Examination. *Teaching and Learning in Medicine* 29(2):115–122

249. Vogels EA. 2023. A majority of Americans have heard of ChatGPT, but few have tried it themselves

250. Weidinger L, Uesato J, Rauh M, Griffin C, Huang PS, et al. 2022. *Taxonomy of risks posed by language models.* In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229

251. Veinot TC, Mitchell H, Ancker JS. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association: JAMIA* 25(8):1080–1088

252. Smith B, Magnani JW. 2019. New technologies, new disparities: The intersection of electronic health and digital health literacy. *International Journal of Cardiology* 292:280–282Publisher: Elsevier

253. Long D, Magerko B. 2020. *What is AI Literacy? Competencies and Design Considerations.* In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pp. 1–16. New York, NY, USA: Association for Computing Machinery

254. Rodriguez JA, Alsentzer E, Bates DW. 2024. Leveraging large language models to foster equity

in healthcare. *Journal of the American Medical Informatics Association* :ocae055

255. Kim JY, Hasan A, Kellogg KC, Ratliff W, Murray SG, et al. 2024. Development and preliminary testing of Health Equity Across the AI Lifecycle (HEAAL): A framework for healthcare delivery organizations to mitigate the risk of AI solutions worsening health inequities. *PLOS Digital Health* 3(5):e0000390

256. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. ArXiv:2009.13081 [cs]

257. Pal A, Umapathi LK, Sankarasubbu M. 2022. MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. ArXiv:2203.14371 [cs]

258. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. ArXiv:1909.06146 [cs, q-bio]

259. Lai VD, Ngo NT, Veyseh APB, Man H, Dernoncourt F, et al. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. ArXiv:2304.05613 [cs]

260. Deas N, Grieser J, Kleiner S, Patton D, Turcan E, McKeown K. 2023. Evaluation of African American Language Bias in Natural Language Generation. ArXiv:2305.14291 [cs]

261. Ghosh S, Caliskan A. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. *arXiv preprint arXiv:2305.10510*

262. Abacha AB, Mrabet Y, Sharp M, Goodwin TR, Shooshan SE, Demner-Fushman D. 2019. Bridging the Gap Between Consumers' Medication Questions and Trusted Answers. *Studies in Health Technology and Informatics* 264:25–29

263. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine*

264. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, et al. 2023. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *medRxiv* :2023–02

265. Ayers JW, Zhu Z, Poliak A, Leas EC, Dredze M, et al. 2023. Evaluating artificial intelligence responses to public health questions. *JAMA Network Open* 6(6):e2317517–e2317517

266. Li SS, Balachandran V, Feng S, Ilgen J, Pierson E, et al. 2024. MEDIQ: Question-Asking LLMs for Adaptive and Reliable Clinical Reasoning

267. Antoniak M, Naik A, Alvarado CS, Wang LL, Chen IY. 2024. *NLP for Maternal Healthcare: Perspectives and Guiding Principles in the Age of LLMs*. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1446–1463

268. Cestonaro C, Delicati A, Marcante B, Caenazzo L, Tozzo P. 2023. Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. *Frontiers in Medicine* 10:1305756

269. Bastani H, Bastani O, Sungu A, Ge H, Kabakcı O, Mariman R. 2024. Generative AI Can Harm Learning. *Available at SSRN* 4895486

270. Vicente L, Matute H. 2023. Humans inherit artificial intelligence biases. *Scientific Reports* 13(1):15737

271. Sung JJ, Poon NC. 2020. Artificial intelligence in gastroenterology: where are we heading? *Frontiers of medicine* 14(4):511–517